1. Team Details

Team Name: GenreBridge

Section: B

Team Members :

S.No	Name	Roll Number	Email ID
1	Narendra Sirvi	24bcs10225	narendra.24bcs10225@sst.scaler.com
2	Pujan Khunt	24bcs10138	pujan.24bcs10138@sst.scaler.com
3	Jenil Langaliya	24bcs10323	langaliya.24bcs10323@sst.scaler.com
4	Smit Rupani	24bcs10256	smit.24bcs10256@sst.scaler.com
5	Ritesh Prajapati	24bcs10088	ritesh.24bcs10088@sst.scaler.com
6	Jash Patel	24bcs10632	jash.24bcs10632@sst.scaler.com

Team Leader Name & Email ID: narendra.24bcs10225@sst.scaler.com

---

Cross-Domain Recommendation of Books Using User Movie-Watching Behavior

---

3. Problem Statement

3.1 Background & Context

In the digital content ecosystem, users interact with multiple forms of media such as movies, books, music, and podcasts. Streaming platforms (movies) and reading platforms (books) often operate in isolation, leading to limited personalization when user interaction data is sparse or unavailable in one domain.

Who faces this problem?
Content recommendation platforms, digital libraries, e-commerce platforms (e.g., Amazon, Goodreads), and users seeking personalized recommendations.

Industry/Domain:
Media & Entertainment, Recommender Systems, Personalization, E-commerce.

Why is this problem important?

- Cold-start problem in recommendation systems
- Poor personalization across platforms
- Missed opportunities for cross-selling and user engagement
  Cross-domain recommendations can improve user experience even when direct interaction data is limited.

3.2 Core Problem Definition

The goal of this project is to **recommend books to users based on their movie-watching history**, without requiring explicit book ratings from the user. The system aims to learn latent preferences from the movie domain and transfer them to the book domain using content similarity and representation learning.

Specifically, the project attempts to discover **semantic and thematic relationships between movies and books** (e.g., genre, themes, tone) and use these relationships to generate personalized book recommendations.

Supervised learning is not suitable because:

- Explicit labeled data mapping movies to books or user preferences to books is not available.
- User-book interaction data is sparse or nonexistent.
- The task relies on discovering latent structures rather than predicting labeled outcomes.

---

4. Dataset Description

4.1 Dataset Source - Suggest to get data from genuine sources

Dataset 1: Movie Dataset

- Dataset name: *MovieLens Latest Small Dataset*
- Source: Kaggle
- Source link: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset
- Public: Yes

Dataset 2: Book Dataset

- Dataset name: *Goodreads Books Dataset*
- Source: Kaggle
- Source link: https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks
- Public: Yes

Both datasets are **self-sourced from real-world public platforms** and analyzed independently before integration.

4.2 Dataset Overview (Tentative if not decided yet)

Attribute	Details
Number of rows	Movies: ~45,000, Books: ~10,000
Number of columns	Movies: ~24, Books: ~12
Type of data	Categorical, Text, User-Item
Time period	Movies: 1900–2017, Books: Various
Missing values	Yes (ratings, descriptions, genres)

4.3 Why This Dataset Fits the Problem

Both datasets contain **rich metadata** such as genres, descriptions, tags, and ratings.
Enables **content-based and representation-based similarity learning**.
Allows discovery of cross-domain patterns such as:

- Genre overlap (e.g., Sci-Fi movies → Sci-Fi books)
- Thematic similarity (e.g., dystopian narratives)
- User preference embeddings

---

5. Chosen Advanced ML Technique(s)

Primary Technique (Mandatory)

Recommender Systems

Supporting Technique

Unsupervised Learning

Why This Technique Is Appropriate

- Recommender systems naturally fit personalization and suggestion problems.

- Unsupervised techniques (e.g., clustering, embeddings, topic modeling) help extract latent features without labeled data.

- Techniques such as:
  - TF-IDF / Sentence embeddings on descriptions
  - Cosine similarity
  - Matrix factorization or latent factor models enable cross-domain knowledge transfer.

Expected Output:
A ranked list of book recommendations for each user based on their movie-watching behavior.

---

6. Expected Outcomes & Insights

- **Personalized book recommendations** for users with only movie interaction history.

- Discovery of:
  - Cross-media thematic clusters
  - Latent user preference representations

- Business impact:
    - Improved engagement
    - Better cold-start handling
    - Cross-platform recommendation capability

Success Metrics (beyond accuracy):

- Recommendation diversity
- Coverage
- Qualitative relevance (manual inspection)
- User-profile consistency

---

7. Assumptions & Limitations

Assumptions

- Movie preferences reflect broader narrative and thematic interests.
- Metadata quality is sufficient to capture semantic similarity.
- Users prefer similar themes across different media formats.

Limitations

- No explicit user-book feedback for validation.
- Subjective nature of recommendations.
- Metadata sparsity or noise.
- Cultural and language bias in datasets.

---

8. Project Scope Confirmation

☑ We confirm that the dataset is self-sourced
☑ We understand that the problem statement and dataset **cannot be changed after approval**
☑ We understand that late submissions will not be accepted
☐ Team size is exactly 5 members

---

9. Declaration

We confirm that:

- This proposal is our original work
- We will start project execution **only after approval**
- We will follow all project guidelines strictly